IBM Capstone

CLUSTERING THE CITY OF CALGARY

Choosing where to live

Introduction

The city of Calgary, affectionately known as Cowtown is one of the largest cities in Canada and has a population of about 1.51 million. This city has grown over the years in both population and diversity. It was ranked 5th most livable city in the world for 2019 by The Economist Intelligence Unit. And with the pull of the Rockies, it is hard to disagree.

As with any city, Calgary is like a living organism with many parts and different behaviors. And as a resident of this beautiful city, I find that each neighborhood has a different feel to it. Having lived here for a few years now, I still find it difficult to pick where to live in the city as my life evolves. Many times, I have resulted to reading blog posts such as Best Places to Live in Calgary. However, this does not tackle the issue of fit. In fact, I have lived in a neighborhood that I did not like despite it being on the list. Armed with my new data science skills, I got thinking, what if we could predict what communities or neighborhoods would be a fit.

Now, my goal is to segment the city of Calgary to find the commonalities within the communities.

Data requirement

Following the thought process of how people decide where to live, I would determine the basis for gathering data. Some of these criteria include

1. Cost of property
2. The amenities in the area - which may include parks, schools, lakes, gyms, grocery, etc
3. The crime rate in the area

4. Demography of the people living there
5. Distance to their most frequent destination - typically the workplace

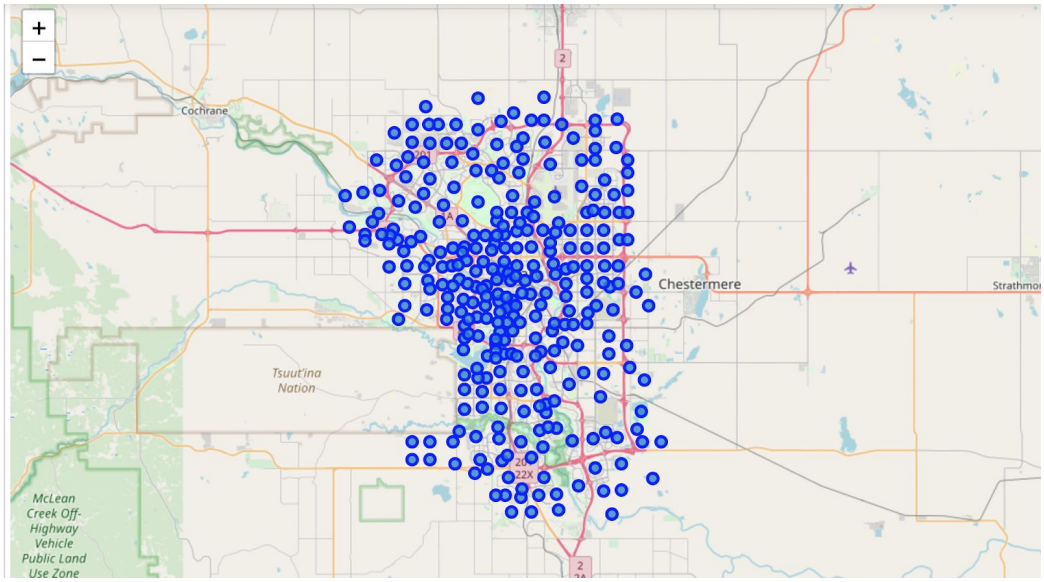
Now that we have an idea of what most people look for while considering what neighborhood to live in, we can gather some data. Below is the list of datasets I would be using for our analysis and model.

- Demographic data: This was gotten from the census data 2019 that can be found on Calgary Open Data. Using this data for each community, we would have data containing information on population, gender, if a community is built out or developing, number of people renting vs own a property, number of people in certain age groups.
- Property assessment: One important factor for picking a community to live in is the cost of properties. Knowing it would be more difficult to scrape actual sale prices from MLS which signal the actual market value, I would use the assessment values provided by the city of Calgary. Although it is less accurate than the actual market value, it is the estimation used to determine property taxes and it provides a more complete dataset of properties in the city. This dataset includes residential and non-residential property values. However, for our purposes, we would focus on residential values only. You can find this data here (Calgary Open Data)
- Venue data: To understand the amenities around the community including shops, grocery stores, cafes etc, I would use Foursquare's API to pull venue data and use it in the model.
- Community boundaries data: To display the boundaries of each community on a map, we need the boundary data also from Calgary Open Data.
- Crime data: Also sourced from the Calgary Open Data, we have crime data for each community for a few years. We can use this to determine how much crime occurs in an area and the severity of the crime.

Methodology

Having gathered all the necessary data needed, it was time to pre-process and analyze them to see what would be useful. First was visualizing the geographic details of the 306

communities in Calgary using Folium. This included the industrial parks and communities that are yet to be built.



Next was examining the property value data. The data includes the addresses, property value, class of properties - residential or non-residential, the community and location data.

	ROLL_YEAR	ROLL_NUMBER	ADDRESS	ASSESSED_VALUE	ASSESSMENT_CLASS	ASSESSMENT_CLASS_DESCRIPTION	RE_ASSESSED_VALUE
0	2019	201443785	4448 FRONT ST SE	1.482170e+09	NR	Non-residential	NaN
1	2019	202215166	6455 MACLEOD TR SW	1.310770e+09	NR	Non-residential	NaN
2	2019	201485596	2000 AIRPORT RD NE	8.401600e+08	NR	Non-residential	NaN
3	2019	79120903	1410 OLYMPIC WY SE	7.856400e+08	NR	Non-residential	NaN
4	2019	202270377	10 SHEPARD ENERGY DR SE	7.763626e+08	NR	Non-residential	NaN

For the purpose of this project, the properties were limited to properties with a class value of residential then unnecessary columns dropped. Prior to grouping the data by each community, the results from a check showed that there was only 1 property value for each address. This means buildings such as apartments/condos with multiple owners have their values all summed up in a single value. This created difficulty in determining the average value of properties in each community. With this new challenge, a good proxy for the

average property values was needed. Leveraging data points from the census data, 2 proxies were considered - the Value per Resident in the community and Value per Dwelling in the community.

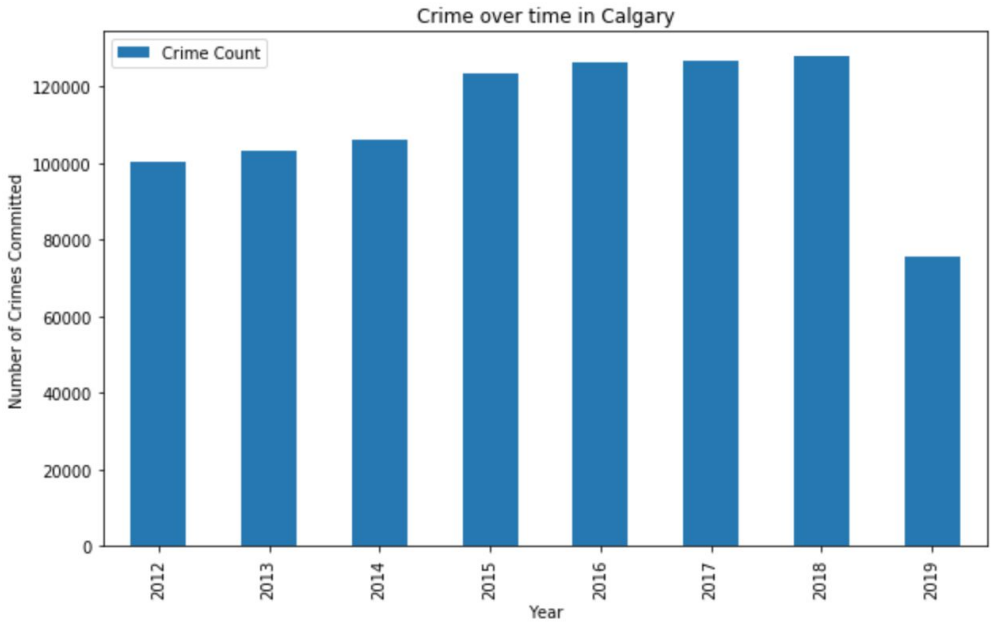
	ASSESSED_VALUE	RESIDENT_COUNT	DWELLING_COUNT	ASSESSED VALUE PER RESIDENT	AVG ASSESSED VALUE
count	2.040000e+02	204.000000	204.000000	2.040000e+02	2.040000e+02
mean	1.071118e+09	6246.500000	2532.392157	2.326416e+05	5.894054e+05
std	8.103312e+08	5042.251491	2015.780770	2.898142e+05	1.184904e+06
min	3.758070e+06	14.000000	5.000000	6.151669e+04	1.155433e+05
25%	5.063902e+08	2345.500000	1019.750000	1.403082e+05	3.281944e+05
50%	8.793782e+08	5470.000000	2418.000000	1.801299e+05	4.340525e+05
75%	1.448186e+09	8680.000000	3475.250000	2.393672e+05	5.513868e+05
max	4.995525e+09	25710.000000	18308.000000	3.811893e+06	1.617889e+07

After reviewing the descriptive analysis data, the mean for the Value per Resident was $232,641 while the Value for Dwelling was $589,405. Comparing these figures to those of other sources tracking the average home prices in Calgary, $451,544, the Value per Dwelling looked like a good average for our purposes. Two other reasons support this choice. First, it is generally accepted that the cities valuations are usually higher than the prices when the property goes on the market. Second, the median of $434,052 in the dataset was close to the median of $487,399 from other sources.

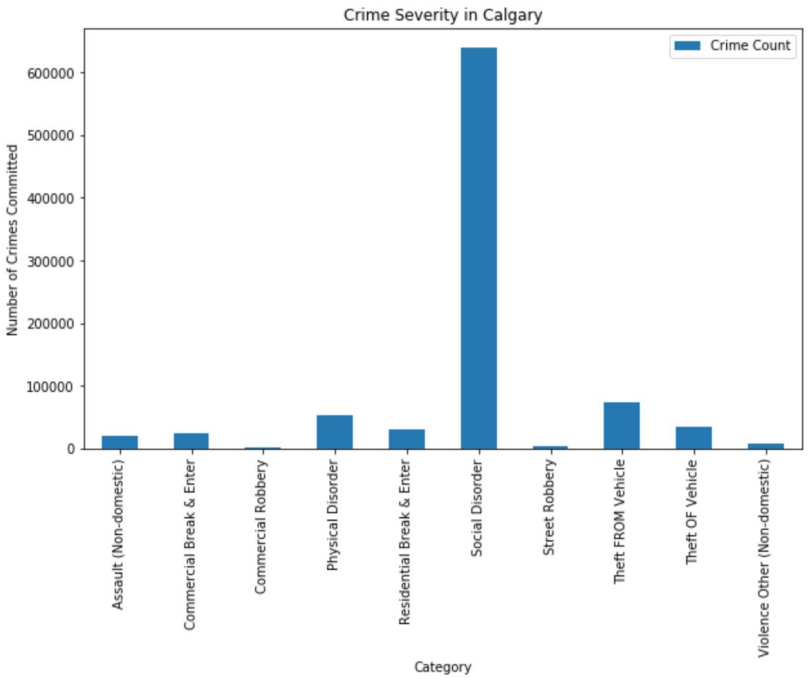
The census data was a goldmine for potential features for the algorithm. Upon examination, it was quick to realize that the useful part of this dataset was the age demographics, the types of homes and if a community was built out or not. The next consideration was what to do with the industrial parks with residents and growth status value of NA. With very few communities in these communities and a total of 956 residents, it was safe to drop them altogether. Rather than using the counts, it determined that a better measure to use was ratios for both the age demographics and home types. The resulting data looked like this.

COMMUNITY_NAME	GROWTH_STATUS	SINGLE_FAMILY	DUPLEX	MULTI_PLEX	APARTMENT	TOWNHOUSE	AGE_0-4	AGE_5-14	AGE_15-19	AGE_20-24	AGE_25-34
LEGACY	1	0.390817	0.128706	0.005061	0.362979	0.112437	0.106854	0.118692	0.050467	0.065109	0.236449
HIGHLAND PARK	0	0.361440	0.176548	0.019763	0.228371	0.010540	0.073476	0.083898	0.023710	0.056540	0.202449
CORNERSTONE	1	0.463035	0.205447	0.053696	0.170428	0.099611	0.081949	0.125378	0.048338	0.047205	0.229230
MONTGOMERY	0	0.490313	0.267760	0.004968	0.065574	0.055142	0.063566	0.085493	0.037431	0.070210	0.222148
TEMPLE	0	0.617198	0.163140	0.000268	0.031342	0.113582	0.061310	0.143937	0.062494	0.063952	0.147217

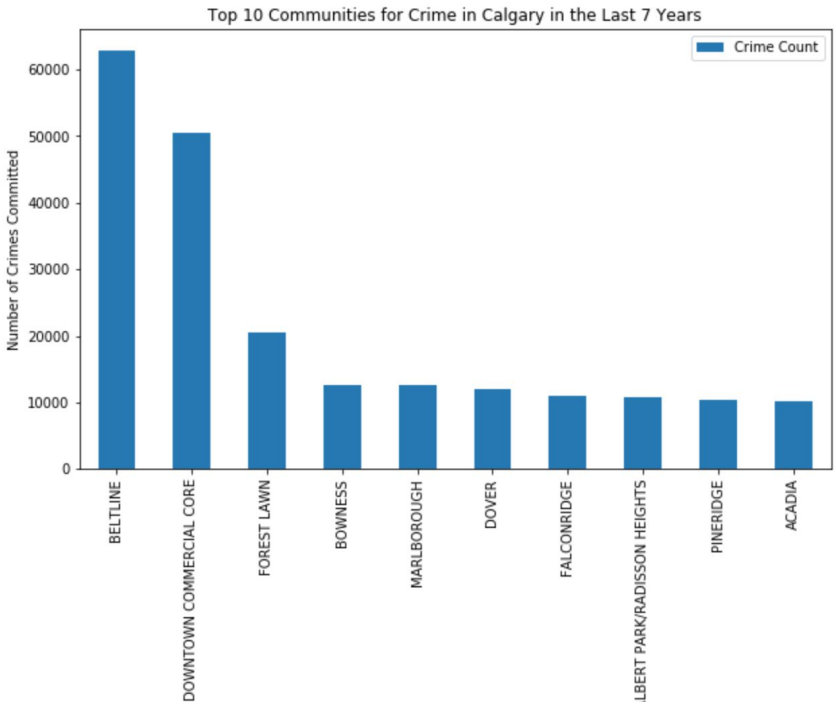
Another parameter that was considered was the amount of crime in the city. The first thing I was curious about was the trend of crime in Calgary. There was no surprise here with a growing population. The number of crimes increased slightly year over year with the exception of 2014 and 2015 with a huge jump. This coincides with the time period of the downturn due to the massive drop in oil prices.



Another thing to note was that the crimes in Calgary are largely skewed towards social disorder.



Therefore, this parameter was dropped from being considered as a feature. The last thing was to see where most of the crimes were committed. The Beltline and Downtown core are where the majority of crimes are committed.



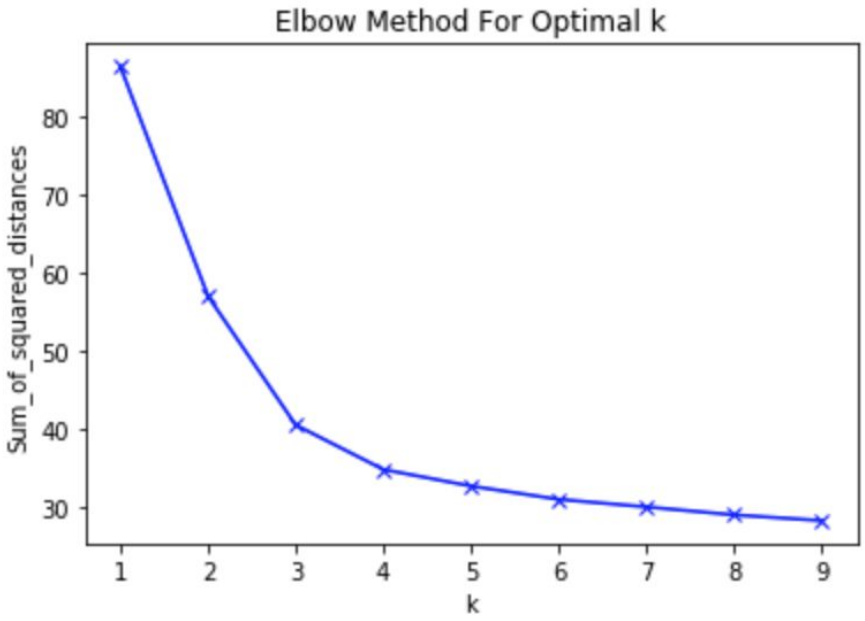
Finally, with a limit of 100 venues and a radius of 1km, the search for the venues in the communities began using Foursquare's API. There are some limitations to the data we get back. For example, the limit of 100 venues which is the max by Foursquare on the current package I'm using, the points selected for the location data may not be central and, the size of the community may be larger or smaller than the selected radius. This means we may not have a complete picture of all the venues in a community. However, the information is sufficient for this project. In total, we got back 283 unique categories. These were grouped to get the Top 10 venues in each community by encoding the data then grouping by the mean frequency for each category.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	01B	Coffee Shop	Restaurant	Golf Course	Event Space	Fast Food Restaurant	Zoo Exhibit	Football Stadium	Food	Food & Drink Shop	Food Court
1	01C	Campground	Convenience Store	Soccer Field	Theater	Garden	Furniture / Home Store	Frozen Yogurt Shop	Garden Center	Fried Chicken Joint	Fast Food Restaurant
2	01F	Construction & Landscaping	Golf Course	Zoo Exhibit	Food	Food & Drink Shop	Food Court	Food Truck	Football Stadium	French Restaurant	Fried Chicken Joint
3	01H	Playground	Zoo Exhibit	Farmers Market	Filipino Restaurant	Food	Food & Drink Shop	Food Court	Food Truck	Football Stadium	French Restaurant
4	01I	Playground	Home Service	Zoo Exhibit	Farmers Market	Filipino Restaurant	Food	Food & Drink Shop	Food Court	Food Truck	Football Stadium

Modeling the data

Following the pre-processing and exploration of our dataset has been explored, it was time for modeling the data. I opted for using the unsupervised learning algorithm, K-Means. Also, it is one of the most common algorithms for clustering.

A loop was written to train the algorithm using different k values to find the optimal number of clusters (k). The Elbow Method was chosen, comparing the sum of squared distance with the k value to find what point we see a minimal gradient change.



Although it could be argued that the optimal k is 3 or 4, I chose 4 since it is the last point before there is minimal change in the gradient.

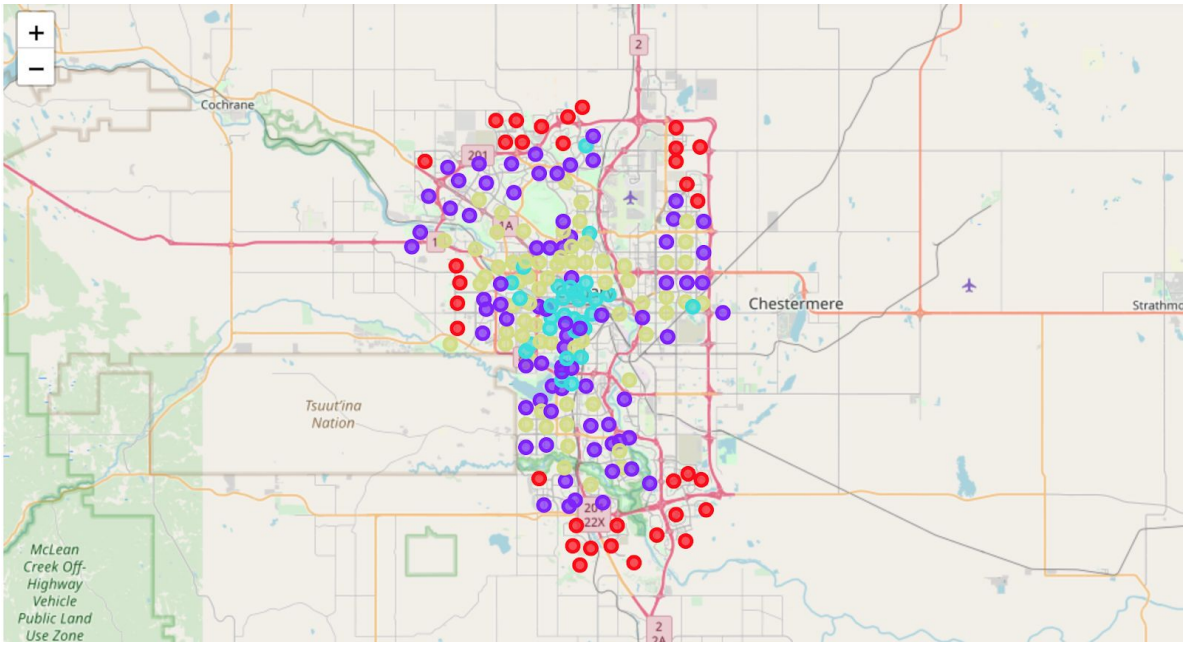
Results

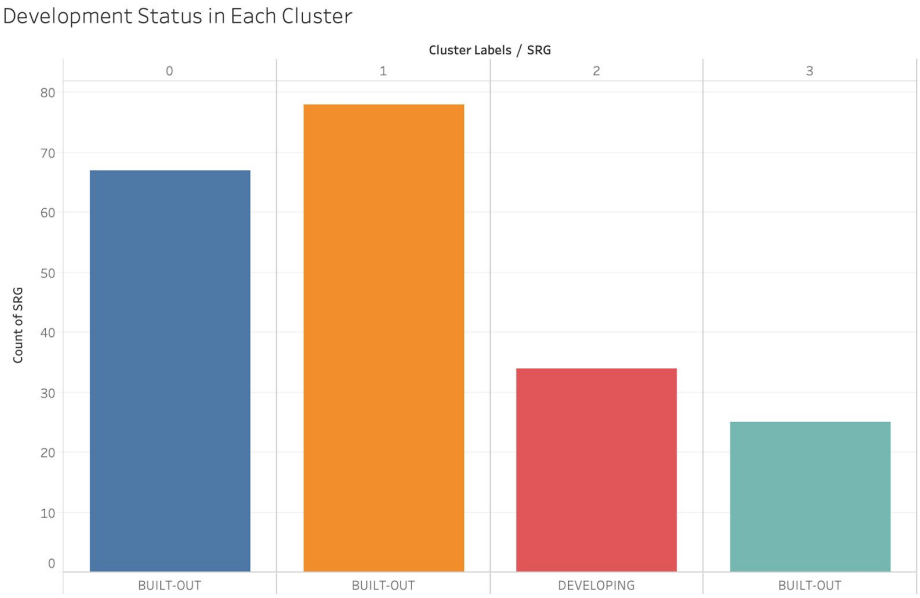
Having trained the model with 4 clusters, the data returned looked like this.

	COMMUNITY_NAME	COMMUNITY_CODE	Cluster Labels	GROWTH_STATUS	SINGLE_FAMILY	DUPLEX	MULTI_PLEX	APARTMENT	TOWNHOUSE	AGE_0-4	...	1C C
0	ABBEYDALE	ABB	1	0	0.685355	0.140164	0.000000	0.000000	0.111165	0.061440	...	
1	ACADIA	ACA	3	0	0.452239	0.026159	0.000198	0.401506	0.076100	0.051236	...	Re
2	ALBERT PARK/RADISSON HEIGHTS	ALB	3	0	0.301993	0.157475	0.008970	0.263787	0.114950	0.082464	...	
3	ALTADORE	ALT	3	0	0.416134	0.192694	0.008219	0.190259	0.140639	0.076491	...	
4	APPLEWOOD PARK	APP	3	0	0.567702	0.008282	0.000414	0.332919	0.054658	0.061739	...	

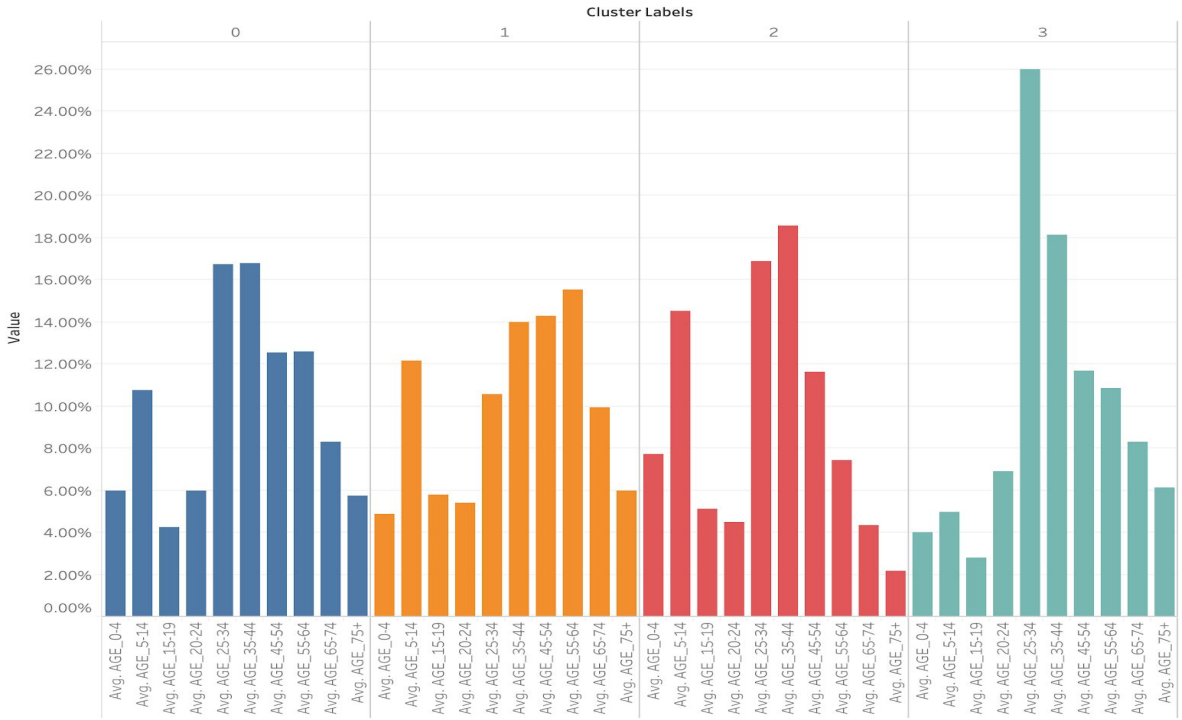
68 communities were in cluster 0, 78 in cluster 1, 34 in cluster 2 and 25 in cluster 3. All of the communities in cluster 2 are located at the edges and are still being developed as

shown in the map and bar chart. Communities in cluster 3 are mostly in the inner city area while cluster 0 and 1 take up the space in between.

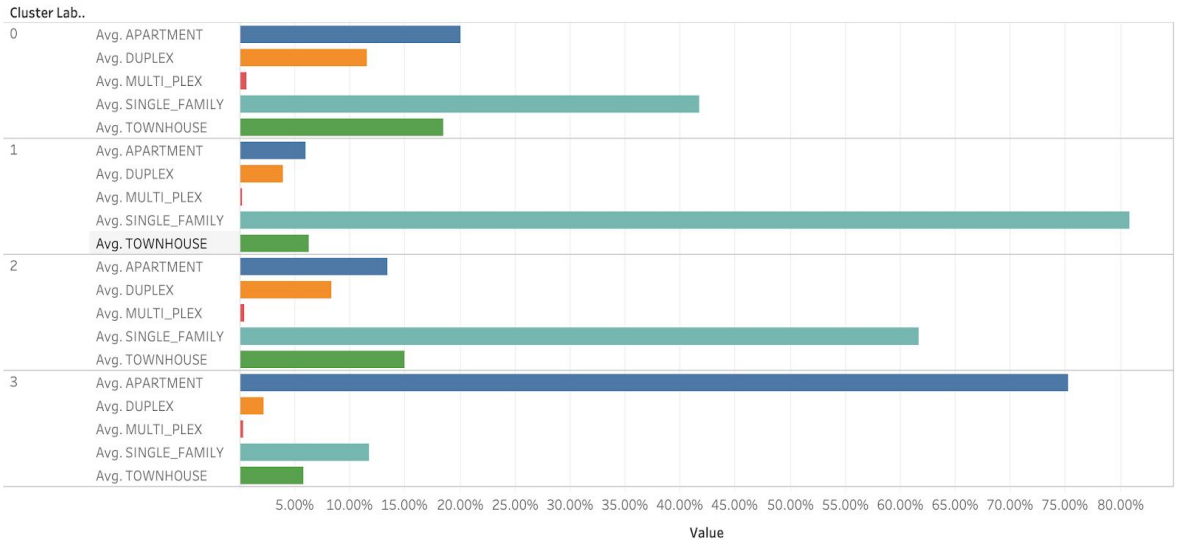




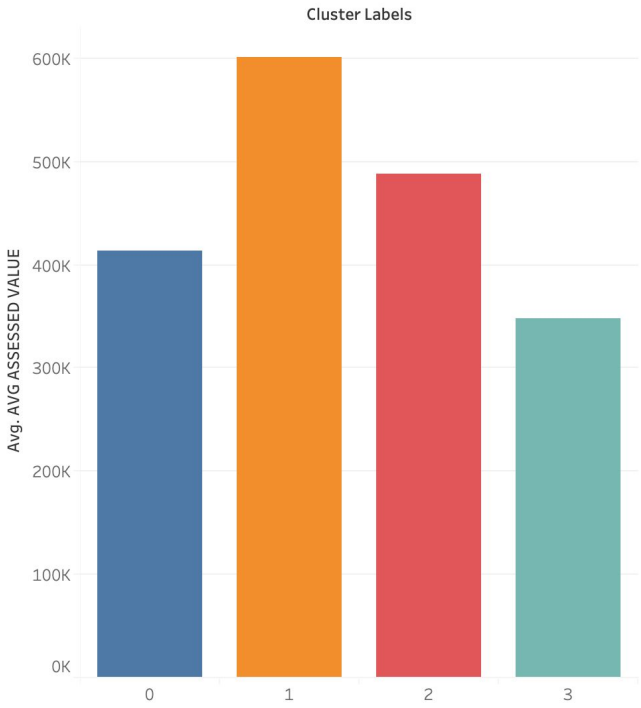
When the age groups that live in each cluster are considered, we notice a few things. First, cluster 2 has a younger population - the majority being in the ages 25-44 at 35.45% of the population when combined. We also see a huge part of the population being children. Similar patterns in cluster 2 are observed in clusters 0 and 1 when it comes to the children but this is where the similarities end. Cluster 1 has a much older population with age 55-64 leading the charge at 15.52% of the population. Cluster 0 also has a young population, however, there is a stronger presence of older people living in those communities when compared to cluster 2. Finally, communities in cluster 3 are mainly populated by young people especially the age group 25-34 at about a quarter of the population. Also, there are fewer kids at 3.99% compared to 5%+ in all the other clusters.



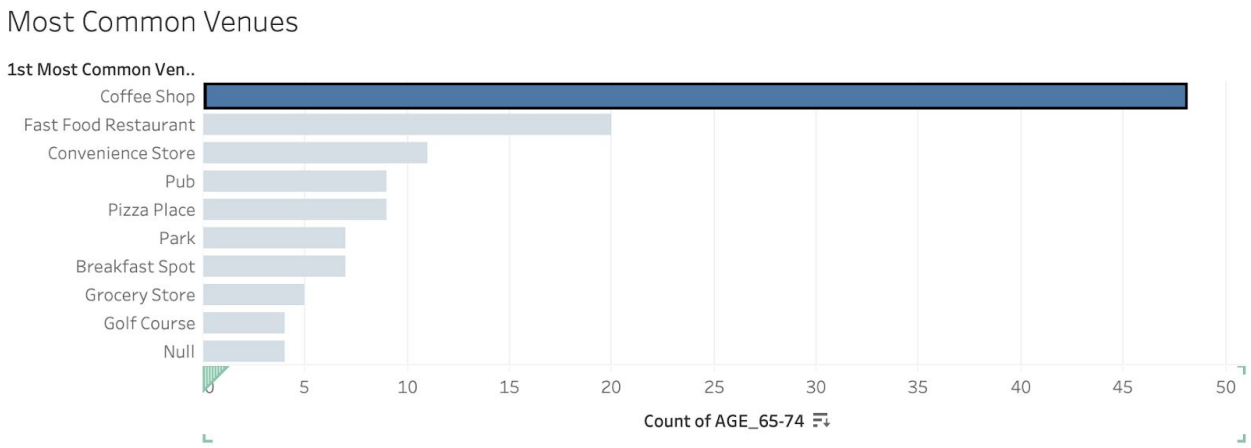
The core focus for the communities in Calgary is on single-family homes except for the communities in cluster 3.

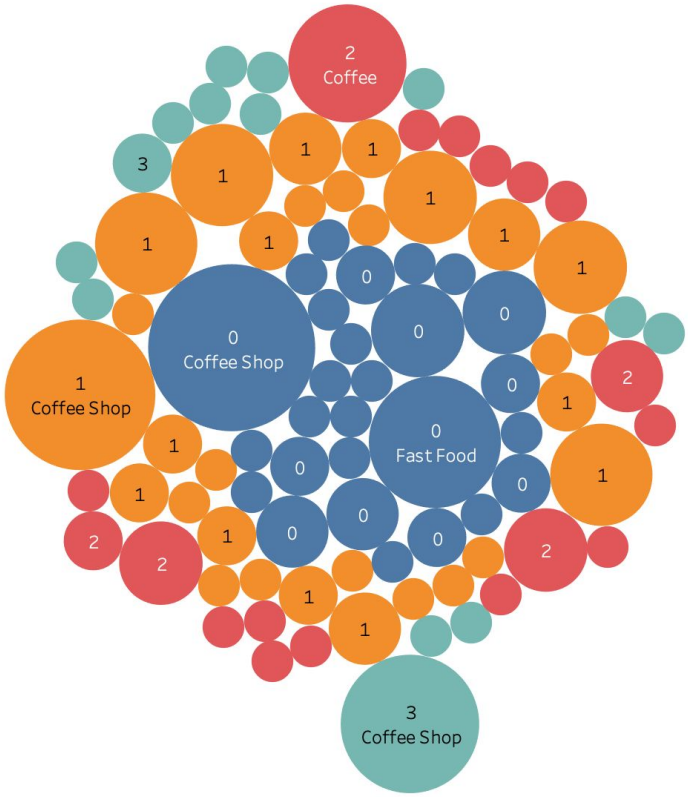


Communities in cluster 1 on average are more expensive than the rest while those in cluster 3 are the least expensive on average. The low average value for cluster 3 could be due to the dominance of apartments in the communities. Furthermore, there is no consistent distribution of property values between the clusters. However, when we examine the property values based on the sectors in the city, we notice the center is the most valuable while the east is the least.



Coffee shops are prevalent in Calgary, being the most common venue in all the clusters. The communities in cluster 3 are dominated by coffee shops and pubs. In cluster 0, coffee shops prevail, then fast-food restaurants, pizza places, and parks. Cluster 1 has more variety in what is on offer including coffee shops, parks, food places, convenience stores. Finally, cluster 2 leads with coffee shops, restaurants, and pharmacies.





Discussion

This evaluation of the city of Calgary shows that If you have a young family, you want to consider living in communities in cluster 2 while those without kids will be well suited to live in communities in cluster 3. Based on the characteristics of the clusters, the categories are characterized as follows:

1. Cluster 0: These communities are similar to those in cluster 2 with a slight variation, more townhouses, and apartments. With more options in types of accommodation available, these tend towards a slightly younger population than cluster 2. Therefore I'll call it The Blend.
2. Cluster 1: These communities are also family-oriented. However, they are accommodated by families that are older in age. Also, when we look at where they are located on the map, they represent communities that were once at the edge of the city before the city expanded. Therefore, I'll call this cluster The Mature.
3. Cluster 2: These communities are family-oriented and focused on younger families. This could be driven by price when a family expands and moves from more central locations to get space in a new build property. Therefore I'd call this cluster The Builder.
4. Cluster 3: Dominated by apartment buildings and a young population, these communities are located around the heart of the city. I'll call this The Lively.

Conclusion

In the beginning, I started with wanting to know where in the city of Calgary a person would like to live in by grouping similar communities together. This has been achieved by using an unsupervised machine learning algorithm, K-Means to create clusters of communities that are similar. The demographics of the population, nearby venues, the types of housing and the development status of the community were taken into account to produce 4 clusters of communities. Young people with no kids tend to live close to the city center with lots of apartments while families live in other parts of the city. The older generation (65+) tends to live in communities other than those at the edges.

Using this work, a recommendation can be made to a person based on criteria that fit the parameters chosen to narrow down communities they may be interested in living in. Then the selection may be narrowed further using the average assessed value.